Preliminary Cross-cultural Evaluation of Expressiveness in Synthetic Faces

Jonas Beskow[1], Loredana Cerrato[1], Piero Cosi[3], Erica Costantini[2], Magnus Nordstrand[2], Fabio Pianesi[2], Michela Prete[4], Gunilla Svanfeldt[1]

[1] KTH Speech Music and Hearing, 10044 Stockholm - Sweden
{beskow, loce, magnusn, gunillas}@ speech.kth.se
[2] ITC-irst, 38050 Povo (TN) - Italy
{costante, pianesi}@itc.it
[3] ISTC-SPFD CNR 35121 Padova -Italy
cosi@csrf.pd.cnr.it
[4] Department of Psychology, University of Trieste, 34134 Trieste – Italy
prete@psico.units.it

Abstract. This paper reports the results of a preliminary cross-evaluation experiment run in the framework of the European research project PF-Star[1], with the double aim of evaluating the possibility of exchanging FAP data between the involved sites and assessing the adequacy of the emotional facial gestures performed by talking heads. The results provide initial insights in the way people belonging to various cultures react to natural and synthetic facial expressions produced in different cultural settings, and in the potentials and limits of FAP data exchange.

1 Introduction

The evaluation of cross-cultural similarities and differences in the production and recognition of facial expression is becoming an important issue in the field of multimodal and multilingual communication research.

Within the European project Pf-Star a whole work package is dedicated to synthesis of facial expressions of emotions, and much effort has been spent to design a methodology for evaluation of the emotional facial gestures performed by 3D animated talking heads. Particular attention has been paid to harmonize the infrastructure and facilitate the exchange of models and data between the involved sites. Exchanging data not only increases integration and cooperation possibilities, but also gives each data set a higher productivity, since it can be used at different sites, and the results compared, even in a cross-cultural perspective.

The experiment we describe in this paper aimed at identifying crucial areas of synchronization of data and methods across sites, addressing cultural issues in the recognition of emotional expressions from synthetic agents. The experiment involved

[1] Project website: http://pfstar.itc.it

the Swedish and Italian partners. The design principles were inspired by Ahlberg, Pandzic and You's [1] evaluation procedure for MPEG-4 facial animation players. They propose to measure the expressiveness of a synthetic face through the accuracy rate of human observers who recognise the facial expression, and to compare the expressions of the synthetic face with those of the "original" human face, upon which they are based. Similarities and dissimilarities between their methodology and ours are discussed in details in [2].

Section 2 of the paper illustrates how the test data were acquired and exchanged across sites. Then in section 3 the experiment is described and some preliminary results are presented and discussed in section 4.

2 Materials

Preparation of data involved: recording actors uttering a series of stimuli acted with emotions in different conditions, production and exchange of the related MPEG-4 FAP (Facial Animation Parameters) files, and animation of the FAPs sequences using different synthetic faces.

Similar data acquisitions have taken place at both the involved sites, using opto-electronic systems able to capture the dynamics of emotional facial expressions with very high precision. The Swedish corpus was recorded using a four-camera Qualysis MacReflex system [3], capturing 35 markers at a rate of 60 frames per second. The Italian corpus was collected with the Elite system [4], which uses two cameras with a frame rate of 100 Hz to capture 28 markers. Both corpora include a common sub-set of data for cross-site comparisons, consisting of 10 nonsense words (which have a very similar pronunciation both in Italian and Swedish) uttered 3 times each with 3 emotions: *neutral, angry* and *happy*. The nonsense words recorded in Sweden were uttered in isolation, while those recorded in Italy were preceded by an opening word and followed by a closing one, which were uttered with neutral expression, and were then cut out of the video files, in order to reduce the differences with the Swedish stimuli. However, some differences could not be cancelled, because the Italian videos were neither starting, nor finishing with a rest position of the mouth. Both the recorded speakers were male actors in their thirties; though the Italian speaker was a professional actor and the Swedish one was an amateur.

Two synthetic 3D face models were used in the study, one originating from Sweden [5] and one from Italy [6]. The Swedish face, a male, is made up of approximately 1,500 polygons, whereas the Italian face is a textured young female built using around 25,000 polygons. Both models adhere to the MPEG-4 Facial Animation (FA) standard, which makes it possible to drive them from the same data. The FAPs are normalized according to the MPEG-4 FA standard, so that they are speaker-independent. The point trajectories obtained from the motion tracking systems described above were converted into FAP streams with custom made software. The FAP streams were then used to animate the synthetic faces.

3 Experiment

One group of Italian (47 volunteer students from the University of Trieste) and one group of Swedish participants (30 volunteers from the University of Stockholm and KTH) were confronted with four blocks of 12 video-files each: 1) Italian actor, 2) Swedish actor, 3) Swedish synthetic face playing both Italian and Swedish FAP-files, and 4) Italian synthetic face playing both Italian and Swedish FAP-files, for a total of 48 stimuli per participant. Two nonsense words, *ABBA* and *ADDA*, uttered with three emotional states (*happy*, *angry* and *neutral*), were selected from the common sub-set of data. The stimuli were played without the audio.

Before the experimental session the participants were given written instructions and were involved in a short training session to familiarise with the task. During the experimental session, the video-files were presented individually on the computer screen, in a randomized order. After each of presented video-file, the participants were asked to choose, on the answering sheet, among the three available labels for the emotional states. At the end of the experimental session, they were also asked to fill in a short questionnaire about their impressions concerning the faces.

4 Results and Discussion

The average percentages of correct recognition, reported in Table 1, show that both human faces got higher rates than synthetic faces. Responses from Italian and Swedish participants have been collapsed, since there were not any significant differences between them (see below).

Table 1. Percentages of correct recognition for each emotion and condition. IT = Italian, SW = Swedish, ACT = actor, SYN = synthetic face.

	IT ACT	SW ACT	IT SYN IT-FAP	IT SYN SW-FAP	SW SYN IT-FAP	SW SYN SW-FAP
Angry	92%	81%	54%	23%	41%	66%
Happy	67%	88%	84%	79%	41%	77%
Neutral	68%	91%	71%	94%	71%	79%
All emotions	**76%**	**87%**	**70%**	**65%**	**51%**	**74%**

We performed two separate loglinear (multinomial logit) analysis [7] of the data with dichotomised responses (correct vs. wrong). In the first the independent variable were the actor (Italian vs. Swedish), the presented emotion, and the subjects (Italian vs. Swedish). The results indicate ($p<.01$): an overall strong tendency towards correct responses; no effects (both main and interactions) of subjects on responses; a significant lowering of recognition rate by the Italian actor on all presented emotion, with the exception of *anger*. In the second loglinear analysis the independent variables were: the subjects, the synthetic face (Italian vs. Swedish), the type of FAP files (Italian vs. Swedish) and the presented emotion. The results indicate ($p<.01$): a) an overall trend towards correct responses; b) that the Swedish FAPs negatively affect

the Italian face on *anger*; c) that the Italian FAPs negatively affect the Swedish face on *happiness*. This suggests than even if exchanging FAPs is technically feasible, one should be careful in assuming that it does not have any consequence. As it turns out, for some reasons FAPs produced in one place are played better by a face produced at the same site. Further experimentation is needed to clarify this point.

As to comparison among the emotional states, there is a significant ($p<.01$) trend for *happiness* to be recognised better than *anger*.

According to the post-session questionnaire, the Swedish actor was considered the easiest to judge (68%), and the Swedish synthetic face as the hardest. The higher rates for naturalness were obtained by the Swedish actor (54%), followed by the Italian one (39%). Finally, the Italian synthetic face has been judged as the most pleasant (45%).

4 Conclusions

The results of this preliminary evaluation show that there are not differences in the way participants, belonging to two different cultural settings (Italian vs. Swedish), react to natural and synthetic facial expressions produced in different cultural settings. Differences emerge as to the provenance of FAPs, though. Drawing clearer conclusions is not possible at this point, since many factors related to cross-sites differences in recording conditions may have affected the results. We are planning more experiments and observations to these purposes. For sure, before exchanging FAP files becomes a standard practice, it is necessary to pay more attention to the intervening factors, by unifying recording conditions.

References

[1] Ahlberg, J., Pandzic, I. S., You, L. 'Evaluationg MPEG-4 Facial Animation Players' In Pandzic, I. S., Forchhimer, R. (eds), 'MPEG-4 Facial Animation: the standard, implementation and applications', 287-291, Wiley & Sons, Chichester, 2002.

[2] Costantini, E., Pianesi, P., Cosi, P. 'Evaluation of Synthetic Faces: Human Recognition of Emotional Facial Displays'. Technical report. ITC-irst. Trento.

[3] www.qualisys.com

[4] Ferrigno G., Pedotti A. ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing. In IEEE - BME-32, 943-950, 1985.

[5] Beskow, J., 2003. Talking heads – models and applications for multimodal speech synthesis. PhD thesis, TMH/KTH.

[6] Cosi P., Fusaro A., Tisato G., LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model, Proceedings of Eurospeech '03, Geneva, Switzerland, Vol. III, 2269-2272.

[7] Agresti A., 2002. Categorical Data Analysis. John Wiley and Sons. New York.